

Improved Hand-Tracking Framework with a Recovery Mechanism

Imran Achmed¹, Isabella M. Venter¹ and Peter Eisert²

Department of Computer Science

University of the Western Cape¹, Private Bag X17, Bellville 7535, South Africa

Tel: +27 21 959 3010, Fax: +27 21 959 3006

and Image Processing Department

Fraunhofer Heinrich Hertz Institute², 10587 Berlin, Germany

Tel: +49 30 31002 614, Fax: +49 30 31002 190

email: {2507311, iventer}@uwc.ac.za¹ and peter.eisert@hhi.fraunhofer.de²

Abstract—Hand-tracking is fundamental to translating sign language to a spoken language. Accurate and reliable sign language translation depends on effective and accurate hand-tracking. This paper proposes an improved hand-tracking framework that includes a tracking recovery algorithm optimising a previous framework to better handle occlusion. It integrates the tracking recovery algorithm to improve the discrimination between hands and the tracking of hands. The framework was evaluated on 30 South African Sign Language phrases that use: a single hand; both hands without occlusion; and both hands with occlusion. Ten individuals in constrained and unconstrained environments performed the gestures. Overall, the proposed framework achieved an average success rate of 91.8% compared to an average success rate of 81.1% using the previous framework. The results show an improved tracking accuracy across all signs in constrained and unconstrained environments.

Index terms: hand-tracking, occlusion handling, Scale Invariant Features Transform (SIFT), sign language recognition

I. INTRODUCTION

In recent years, the importance of communication has been symbolised by the social and socio-economic opportunities it provides [2]. The advancement in mobile technology enables millions of people to benefit from this rich form of social communication and information exchange. Unfortunately, the hearing impaired or Deaf¹, who use sign language as their primary means of communication, are unable to interact socially or convey information with the hearing population [2]. To bridge this communication gap, an automated translation system is required. Such a system is complex and encompasses a multidisciplinary research area that involves natural language processing, linguistics, image processing and artificial intelligence. One of the components of the system is concerned with the recognition of South African Sign Language (SASL) and translating it to English or any other spoken language. The recognition of SASL is challenging due to the complexities involved in the visual interpretation of signed gestures. SASL gestures are collectively represented by facial expressions, hand shapes, hand

movements and hand location. Recognising hand movements and locations fall under the broad term hand-tracking. In SASL, the right and left hands have individual characteristics that convey different meanings. Therefore, to accurately translate from SASL to a spoken language, it is necessary to identify and track each hand independently. When distinguishing between the hands while tracking, three additional challenges should be addressed: (1) dealing with occlusion factors; (2) identifying the right and left hands during and after occlusion has occurred; and (3) recovering from a failure while tracking.

In this paper, a tracking recovery algorithm is proposed that builds on an independent hand-tracking framework presented in our previous research [1], which will be referred here forth as the initial framework. The research involves optimising the initial framework to better handle occlusion. It integrates the tracking recovery algorithm to improve the discrimination between and tracking of the hands. The optimised framework, referred to hereafter as the proposed framework, identifies skin clusters that are likely to be the hands or face using connected components labelling, thereby reducing noisy areas. Each cluster is assigned a unique label to identify a hand as either right or left. These clusters are associated temporally in a non-Bayesian framework and are tracked throughout an image sequence.

When tracking the hands, many strong features exists that links the hand to the arm, clothes, watch or any other object that is in close proximity to the hand. These features are referred to as support features and are collectively used to assign a “confidence” vote to skin clusters identified in an image. The skin clusters with the highest votes are used to automatically identify and relocate the hands associated with their respective support features. Support features that belong to the set of foreground keypoints are given a higher vote than those that belong to the set of background keypoints. Overall, an average tracking success rate of 81.1% and 91.8% was obtained using the initial and proposed framework, respectively.

The rest of the paper is organised as follows: section II discusses the related work; section III presents the optimised framework and integration of a novel tracking recovery algorithm; the experiments and results are analysed in section IV; and section V concludes the paper and proposes future work.

¹ Deaf refers to people that use South African Sign Language as their primary language.

II. RELATED WORK

The process that continuously estimates the hand location and movements throughout an image sequence is referred to as hand-tracking [5]. A number of hand-tracking approaches have been proposed and vary from those using an auxiliary means to those using a purely passive means.

Auxiliary hand-tracking makes use of devices such as data suits, gloves or position markers to measure the spatial positions and joint angles of the hands [11]. Although the hardware used in these approaches usually offer near to real-time performance and more accurate information, it is an impractical and inconvenient solution to sign language recognition. Furthermore, it would require calibrating the equipment to suit each individual's needs.

Passive hand-tracking approaches are able to determine the spatial positions of the hands by using various image processing algorithms in non-invasive ways. These approaches offer more practical solutions and have the capabilities of achieving near to real-time performance.

Roussos et al. [15] proposed a framework for the recognition of sign language videos. They applied skin colour modelling along with morphological filtering to detect and segment the hands. They handled occlusion by tracking the hands and face using a forward-backward prediction based on statistical prior information. They further extracted hand-shape features using affine modelling of hand-shape appearance images to determine the hand pose. Their framework was evaluated on the BU400 dataset² and obtained a sign recognition accuracy of 83% and 82% based on 26 and 40 sign language gestures respectively.

In Liu and Zhang [12], a particle filter framework combined with local binary patterns and colour cues was used to track the hands. They showed that by combining local binary patterns with colour cues, a more robust hand-tracking method can be achieved than with either cue alone. Similarly, Spruyt et al. [16] used a particle filter framework; they however combined it with colour and motion cues to track the hands. They suggest that by combining skin colour, edge detection, colour clustering and motion detection, it increases their framework against illumination invariance. They furthermore suggest that by combining the colour and motion cues in their particle filter framework, their system would automatically recover from failure and would not need an initialisation phase. Their results were visually presented.

The advantage of following a passive approach to sign language recognition compared to auxiliary approaches is that it is inexpensive and has the capabilities of achieving near to real-time performance. This research therefore follows a passive approach as it would be more applicable to hand-tracking in unconstrained environments.

Many researchers, who proposed passive methods to detect and track the hands, do not make provision for a recovery phase in their tracking algorithm. Although Spruyt et al. [16] suggest their particle filter framework automatically recovers from failure, it can be argued that particle filters alone cannot be used as a tracking recovery mechanism, since particle filters largely depend on its likelihood function to make a decision on which object to track. This is further complicated when the hand shape changes.

When attempting to recover from tracking failure, one needs to consider that objects surrounding the tracked object may possess as much information as the tracked object itself. Therefore, instead of explicitly finding the tracked object, the surrounding objects can be used to assist in locating the tracked object. Using surrounding objects is very useful especially in cases where the appearance of the tracked object changes considerably.

Cerman et al. [4] applied this concept to the general object tracking case. They proposed a tracker, based on foreground and background appearance cues, that identifies which image regions move coherently with a tracked object. Their tracker is characterised by an object model, comparison model and the object location. The object model refers to the appearance of the tracked object and the companion model refers to the image regions used to assist tracking where it is adapted on-line in each step of tracking. They suggest the size of the companion model should cover an area larger than the tracked object. They subjectively evaluated their tracker on four video sequences and showed a positive result.

In order to recover from tracking failure, the same concept was used in this research and a novel tracking recovery algorithm is proposed, largely inspired by work of Cerman et al. [4].

III. IMPROVED HAND-TRACKING FRAMEWORK

In the following sub-sections, the optimised framework will be discussed. The discussion will deal with the improved data association of skin identified clusters to better handle occlusion. It will also discuss the tracking recovery algorithm and how it is integrated into the framework.

A. Cluster Selection

In this research, the method to select skin clusters in a frame is similar to the approach discussed in [1]. In order to identify skin-coloured pixels in an image, some researchers employ a trained model [8]. These models rely on the skin-colour range on which it was trained and need to be re-trained if small changes should occur or it would easily fail if large changes should occur. The proposed research method employs a more efficient means to directly identify skin-colour distribution of an individual in an image and adaptively changes the colour distribution throughout an image sequence. The skin-colour distribution is determined by using the area around the nose to determine the skin-colour of an individual in every frame [1]. This ensures that the optimal colour distribution can be extracted without being negatively affected by any eyes, lips or facial hair. By back projecting the colour distribution, the skin identified areas such as the hands and face, would be highlighted. To extract these regions of interest as clusters, connected-components labelling is used.

Connected-components labelling is a sequential two-pass algorithm that assigns a set of pixels into components using the level of its pixel connectivity and thereafter labels each pixel accordingly. The algorithm passes through each two-dimensional (2D) binary image twice, and can use either 4-connectivity or 8-connectivity labelling [6].

This research uses the 8-connectivity labelling mask since connected pixels will be searched for in each direction. In the first pass, the mask moves from the top-left to the

² Boston University American Sign Language dataset.

bottom-right of an image where each skin-coloured pixel is assigned a temporary label based on the values of neighbouring pixels that have been processed. If none of the top-left four neighbouring pixels is a skin-coloured pixel, then the current pixel would be assigned a new label; however, if there is only one neighbouring skin-coloured pixel, then its label is assigned to the current pixel. Furthermore, if a skin-coloured pixel contains two or more neighbouring skin-coloured pixels with different labels, then these neighbouring pixels' labels would be stored as being equivalent. After the first pass, the equivalences are used to determine equivalence classes where each class is assigned a unique label. During the second pass, the label of its corresponding equivalence class would replace each temporary label [1].

After applying the connected-components labelling algorithm, the skin coloured regions of interest are extracted as clusters. This is followed by the analysis of each skin-coloured region, where regions larger than a face or smaller than the fist are discarded. This analysis allows the amount of noise in a frame to be reduced [1].

B. Dealing with Occlusion

Tracking the right and left hands of an individual is a challenging task since the hands are similar and the differences cannot be distinguished easily. Moreover, the colours of the hand and face are almost identical, which further complicates the task. It therefore becomes even more challenging when tracking a hand as the tracking may easily fail when the tracked hands crosses the opposite hand or face.

To deal with the tracking of multiple objects, such as hands, that share similar characteristics, this paper proposes a more effective method compared to the previous method [1].

The proposed method extends the work of Argyros and Lourakis [3]. Their method, to handle and track multiple skin-coloured objects, is based on a static background environment and treats each object as a separate entity. This research extends their method to track multiple objects in unconstrained environments. It identifies each object and distinguishes it from other objects: the right hand is identified and distinguished from the left and the hands are distinguished from the face.

The method operates by associating each skin cluster with an object hypothesis and then associating it with time. The correspondence between each cluster and object is however not necessarily one-to-one. It is assumed that an object may be associated with only one cluster and that a cluster may be associated with one or many objects [3]. It is also assumed that the pixels of a cluster can be approximated by an ellipse, which is valid for objects such as hands [3]. Let N be the numbers of clusters present in a scene at time t and o_i , $1 \leq i \leq N$, be the set of skin pixels that image the i -th object [3]. Furthermore, $h_i = h_i(C_{x_i}, C_{y_i}, \alpha_i, \beta_i, \theta_i)$ denotes the ellipse of an object where (C_{x_i}, C_{y_i}) is its centroid, while α_i, β_i and θ_i is the length of the major and minor axis of the ellipse and its orientation on the image plane respectively [3]. Moreover, let $C = \bigcup_{j=1}^2 b_j$, $O = \bigcup_{i=1}^N o_i$ and $E = \bigcup_{i=1}^N e_i$, denote the union of skin-coloured pixels, object pixels and ellipses respectively. Therefore by associating the ellipses with a cluster across time, multiple clusters can be tracked even when occlusion occurs.

C. Associating hands with object hypothesis

After applying the connected-components labelling algorithm, the skin clusters are identified in a 2D binary image. To identify only the skin clusters that are of interest, such as the hands, a background subtraction algorithm is applied to the image sequence. The background subtraction algorithm is based on a mixture of Gaussians that constantly updates the background model in every frame. This results in a foreground mask. This mask is logically AND-ed with the skin detected image to produce a combined image that only highlights skin clusters that have moved, as seen in Figure 1.



Figure 1: Logically AND-ed motion and skin image to form the motion-skin image.

When associating an object hypothesis or ellipse with a cluster, the distance of a pixel to an ellipse is used to determine if the ellipse belongs to the cluster or not.

The distance, $D(p, h)$, from a point $p = p(x, y)$ to an ellipse $e(C_x, C_y, \alpha, \beta, \theta)$ is defined as follows [3]:

$$D(p, h) = \sqrt{\vec{V}^T * \vec{V}}$$

where

$$\vec{V} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{pmatrix} \frac{x - x_c}{\alpha} \\ \frac{y - y_c}{\beta} \end{pmatrix}$$

If the distance, $D(p, h)$, is less than one, equal to one or greater than one, then the given pixel exists within, on or outside the ellipse respectively. In the initial frame of the image sequence, the two-object hypothesis or ellipses are assigned to the hands, one for the right hand and one for the left. This assumption is valid since individuals begin signing in the neutral pose, with the arms and hands on the side of the body.

The parameters for these initial ellipses are directly derived from the statistics of the distribution of pixels belonging to a cluster, where the center of the ellipse is equal to the center of the cluster and the rest of the parameters are computed from the covariance matrix of the bivariate distribution of the location of the clusters' pixels [3]. Therefore, it can be shown that if the distribution is represented by $\Sigma = \begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{xy} & \sigma_{yy} \end{bmatrix}$ then the rest of the ellipse parameters can be defined by [3]:

$$\alpha = \sqrt{\lambda_1}, \beta = \sqrt{\lambda_2}, \theta = \tan^{-1} \left(\frac{-\sigma_{xy}}{\lambda_1 - \sigma_{yy}} \right)$$

where,

$$\lambda_1 = \frac{\sigma_{xx} + \sigma_{yy} + \Lambda}{2}, \lambda_2 = \frac{\sigma_{xx} + \sigma_{yy} - \Lambda}{2} \text{ and } \Lambda = \sqrt{(\sigma_{xx} - \sigma_{yy})^2 - 4\sigma_{xy}^2}$$

D. Tracking the hands

When tracking the hands, there are two rules that govern the association of a cluster's pixels to an ellipse, as stated by Argyros and Lourakis [3]:

- 1) If a skin-coloured pixel of a cluster is located within an ellipse then that pixel is considered to belong to that ellipse.
- 2) If a skin-coloured pixel is located outside both ellipses, then it is assigned to the ellipse that is closest to it.

To handle cases where an ellipse belongs to more than one cluster, the following third rule is applied [3]:

- 3) If there exists only one cluster that is assigned to an ellipse and, at the same time, not assigned to any other ellipse, then the ellipse is assigned to that cluster. Otherwise the ellipse is assigned to the cluster with which it shares the largest number of skin-coloured pixels.

After assigning the skin-coloured pixels to the ellipses, the parameters for the ellipses are re-estimated based on the statistics of the pixels assigned to them.

E. Predicting hand locations

In order to handle occlusion, pixel data from the third frame onwards are associated and based on the ellipses that have been formed in the previous two frames. Based on the assumption that the immediate past can be used to predict the immediate future, a simple linear rule is used to predict the location of an ellipse at time t , based on their locations at time $t - 1$ and $t - 2$. Therefore, by only using the center point of an ellipse in the previous two frames while keeping all other parameters the same, the location of the current ellipse can be predicted. This can be formally stated as

$$\hat{e}_i = e_i(\widehat{C}_{x_i}, \widehat{C}_{y_i}, \alpha_i, \beta_i, \theta_i)$$

where

$$(\widehat{C}_{x_i}(t), \widehat{C}_{y_i}(t)) = 2C_i(t - 1) - C_i(t - 2)$$

This equation therefore asserts that by keeping all other parameters the same, the predicted ellipse will maintain the same direction and magnitude of translation on the image plane. These parameters are however updated when the skin-coloured pixels in an image are assigned to the predicted ellipse. The updated parameters can therefore be used as a good indication of the size and angle of each in the current frame. Examples of the tracking process output are shown in Figure 2.



Figure 2: The tracking process output.

F. Handling stationary cases

In sign language, gestures are made up of movement-hold sequences. In cases where the hands become stationary (a hold position), it would begin to form part of the background model and therefore not be highlighted in the combined motion-skin image. To deal with such cases, each current ellipse is checked for the number of skin pixels that are located in the ellipse. If the number of skin pixels is less than half of the size of the ellipse, the combined motion-skin image is updated using the parameters of the ellipse in the previous time step. Given these parameters, the distance of each skin pixel in the skin detected image is computed to

determine if it is located in the ellipse. All skin pixels in the skin detected image that exist in the ellipse are then stored in a new image, referred to as updated skin image. This image is logically OR-ed with the motion-skin image to form an updated motion-skin image. Finally the updated motion-skin image is used to update the parameters for the current ellipse and used to predict the ellipse for the next frame.

G. Recovering hand-tracking from failure

Tracking hands in unconstrained environments is a non-trivial task since objects in the background may negatively affect the tracking process. This, in many cases, leads to tracking failure [10]. In order to recover from such failures, a tracking recovery algorithm is proposed. This algorithm is based on the concept that objects surrounding the tracked object may possess as much information about the tracked object as the tracked object itself.

This information can be retrieved from the features of the surrounding objects. In this algorithm, these features are extracted using Scale Invariant Features Transform (SIFT) [13] and matched using the Fast Approximate Nearest Neighbour Search Library (FLANN) [14]. SIFT has been developed to extract highly distinctive invariant features of objects that can later be used to perform reliable matching of the same object between images. These features possess attractive properties such as being invariant to rotation and scaling in images as well as being partially invariant to changes in illumination and camera viewpoints. To match these features, FLANN is used. This library has been developed to automatically select the best nearest neighbour algorithm and parameters for any given dataset using a cross validation approach, thereby minimising the predicted search cost while maintaining a high accuracy.

The proposed algorithm is embedded in the hand-tracking framework and operates as follows. While tracking each hand, the parameters of the hand are used to set a region of interest (ROI) around the hand that is twice the width and height of the hand. This region is estimated to be large enough to contain any significant object(s) that may link its existence to a hand. For each frame, keypoints within this ROI are set and its descriptors are extracted using SIFT. Descriptors for the right and left hand are then stored in two separate databases. This allows features for the right and left hand to be matched separately.

From the second frame onwards in the image sequence, keypoints are set and their descriptors are extracted for each cluster that exists in the motion-skin image. For each of these clusters, the descriptors are matched to the descriptors in the respective database. For every successful match, the cluster to which the descriptor belongs, will be given a vote based on two rules: (1) if the keypoint of the given descriptor belongs to a pixel that exists in the motion image, then the cluster to which the descriptor belongs, will be given a vote equal to two. (2) If the keypoint of the given descriptor does not exist in the motion image, then it will be given a vote equal to one. This voting strategy is based on the fact that objects that move with the hand, support the existence of the hand more than objects that are stationary. For each cluster, the votes for all the descriptors that have successfully been matched are added. Given the total votes for each cluster, the cluster with the most votes is assigned as the respective hand. This process is illustrated in Figure 4.

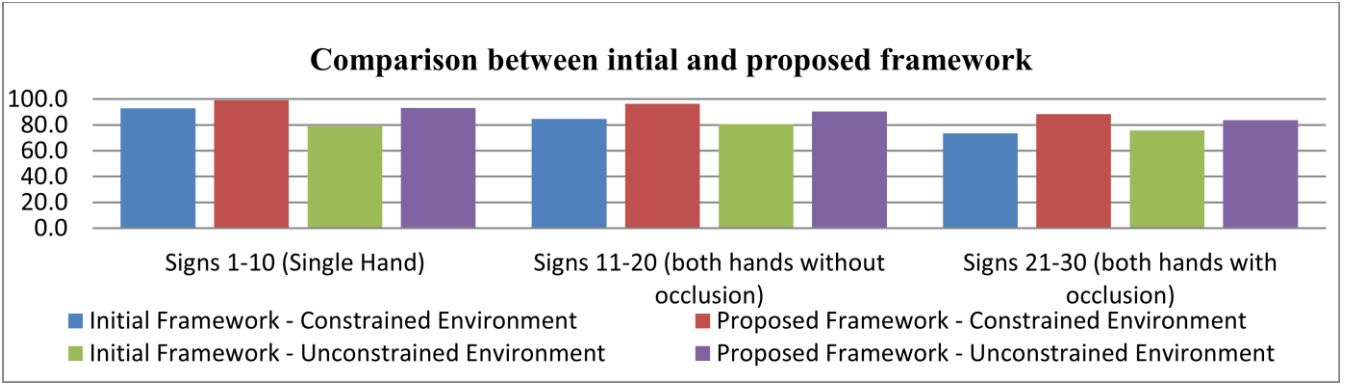


Figure 3: A summary of the comparison between the initial and the proposed framework in constrained and unconstrained environments.

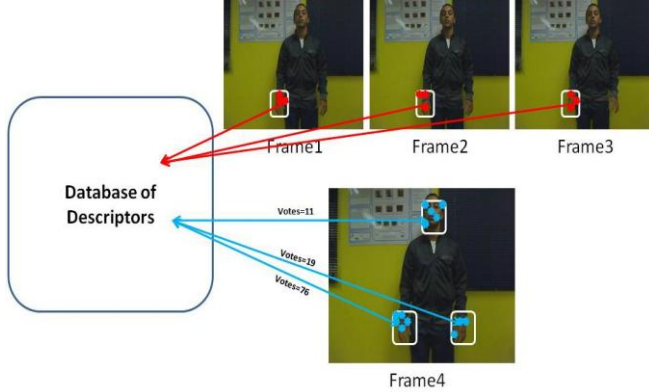


Figure 4: Overview of the algorithm to recover from tracking failure.

IV. EXPERIMENTAL ANALYSIS

This section describes and analyses the experiments used to evaluate the improved hand-tracking framework and shows whether the recovery algorithm assists in recovering from tracking failure. In the experimental setup, a notebook and a single Logitech webcam was used to capture sign language video sequences in constrained and unconstrained environments with varying levels of illumination. These video sequences were captured at approximately 15-20 frames per second with a resolution of 640X480 pixels and an average of 80 frames per video.

The framework was evaluated (as part of a sign language recognition prototype) based on 30 SASL isolated gestures that were each carefully selected from the “*Fulton School for the Deaf SASL Dictionary*” [9]. The selected set of gestures consists of signs that involve the use of a single hand (signs 1 - 10), both hands without occlusion (signs 11 - 20) and both hands with occlusion (signs 21 - 30). Ten individuals with different body types and skin-colour tones, ranging from fair skin-colour tones to very dark tones, performed the SASL gestures. Each individual performed the gestures, in a constrained environment as well as in an unconstrained environment. The constrained environment consisted of a plain static background so that no background objects may interfere with the tracking process. The unconstrained environment consisted of a busy background with several objects of different shapes and colours that may affect the tracking process. The aim of evaluating the framework on different environments was to determine whether objects moving in the background would negatively affect the tracking process. Furthermore, different individuals were used to determine how well the tracking process performs on the different skin colour tones and body types.

In the evaluation of the framework, the tracking process was applied to each video. In each video, the hands are

located next to the body of the signer in the initial frames where the right hand will be on the right side of the body and left hand will be on the left side of the body.

After labelling each hand as either right or left, the hands are tracked through consecutive frames in the image sequence. While tracking the hands, an enclosed red square indicates the hand being tracked is the right hand and an enclosed blue square indicates the hand being tracked is the left hand, as shown in Figure 2.

In the case of occlusion, where the one hand completely covers the opposite hand, the colour of the enclosed square should be either blue or red depending on which hand is in front. In the case of partial occlusion, each hand will be enclosed with their respective coloured square. If either the blue or red square does not surround a hand, then the hand to which the square belongs, is lost. By integrating the recovery algorithm in the tracking framework, it then becomes possible to recover the “lost” hand and continue to track the hands independently.

Subjective evaluation was used for analysis similar to other researchers in this field [7, 12]. After applying the tracking process to each video, the output was analysed by an individual not related to the research. A frame was deemed correct only if both the right and left hands were tracked correctly, i.e. the red or blue square enclosed the right or left hand, respectively. Otherwise, the frame would be labelled as incorrect. To calculate the average tracking success rate per signed gesture, the number of correct frames was divided by the total number of frames in the video. A summary of the tracking success rates for each framework in constrained and unconstrained environments is shown in Figure 3.

In Figure 3, the results indicate the tracking success rate obtained for signs using a single hand (signs 1-10), both hands without occlusion (signs 11-20) and both hands with occlusion (signs 21-30), respectively. The results also show the difference in the average tracking success rate when including the recovery algorithm in the framework. From the results it is seen that in each group of signs, the recovery algorithm has improved the tracking success rate. It also shows that when using the initial framework, each sign obtains an average success rate greater than 60%. However, by using the proposed framework, each sign obtains an average success rate greater than 80% in a constrained environment and 70% in an unconstrained environment.

Furthermore, using the initial framework, the average success rate across all signs is 83.7% and 78.4% in a constrained and unconstrained environment, while using the proposed framework results in an average success rate of 94.6% and 89.0% across all signs in a constrained and unconstrained environment, respectively. These results suggest that in both frameworks, a higher result is obtained

when tracking hands in constrained environments as opposed to unconstrained environments; however, by using the proposed framework, a higher success rate (89.0%) can be achieved in an unconstrained environment when compared to using the initial framework in a constrained environment (83.7%) and an even higher success rate compared to using the initial framework in an unconstrained environment (78.4%). Based on these results, the proposed framework is well suited for unconstrained environments.

When analysing the tracking accuracy according to each individual signer using the proposed framework, the majority of signers obtained an average tracking accuracy greater than 84% with a median of 98.4% across all signs in a constrained environment and a median of 89.6% across all signs in an unconstrained environment. This indicates the proposed framework performed well across the different skin-colour tones and body types. Overall, the initial and proposed framework achieved an average success rate of 81.1% and 91.8%, respectively.

V. CONCLUSION

To provide accurate and reliable hand-tracking, this paper proposed an improved hand-tracking framework that included a tracking recovery algorithm. It involved optimising the framework to increase the effectiveness of dealing with occlusion and integrated the tracking recovery algorithm to recover from tracking failure and to continue distinguishing between the hands. The proposed framework uses connected components analysis to identify skin clusters, which are likely to be the hands or face. These clusters are then assigned unique labels to identify a hand as either right or left. While tracking, many strong features exist that link a hand to an object that is in close proximity to the hand. These features are used to develop a tracking recovery algorithm that helps identify and relocate the hands in cases where tracking failure may occur.

This improved framework was evaluated on 30 SASL phrases performed by ten individuals in constrained and unconstrained environments. Overall, the proposed framework obtained an average tracking success rate of 91.8% compared to an average tracking success rate of 81.1% using the initial framework. The results show that the proposed framework improved the tracking accuracy across all signs in both environments.

To further improve the framework, as future work, an explicitly defined hand-detection algorithm could be integrated to decrease the search space in each frame.

VI. REFERENCES

- 1) I. Achmed, I.M. Venter and P. Eisert, "A framework for independent hand tracking in unconstrained environments", in *Proceedings of the Southern Africa Telecommunication Networks and Applications Conference*, 2012.
- 2) I. Achmed and I.M. Venter, "A discriminative approach to South African sign language recognition from a monocular 2D view", *Proceedings of the Southern Africa Telecommunication Networks and Applications Conference*, East London, 2011.
- 3) A. Argyros, and M. I. A. Lourakis, "Real-time tracking of multiple skin-colored objects with a possibly moving camera," in *Proceedings of the European Conference on Computer Vision*, pp. 368—379, 2004.
- 4) L. Cerman, J. Matas, and V. Hlaváč, "Sputnik Tracker: Having a companion improves robustness of the tracker," in *Image Analysis*, pp. 291—300. Springer Berlin Heidelberg, 2009.
- 5) C.S. Chua, H. Guan and Y.K. Ho, "Model-based 3D hand posture estimation from a single 2D image", in *Image and Vision computing*, vol. 20, no. 3, pp. 191—202, 2002.
- 6) L. Di Stefano and A. Bulgarelli, "A simple and efficient connected components labeling algorithm", in *Proceedings of the International Conference on Image Analysis and Processing*, pp. 322—327, 1999.
- 7) M. Donoser and H. Bischof, "Real time appearance based hand tracking," in *Proceedings of International Conference on Pattern Recognition*, 2008.
- 8) H. Greenspan, J. Goldberger, and I. Eshet, "Mixture model for face-color modeling and segmentation," in *Pattern Recognition Letters*, vol. 22, no. 14, pp. 1525—1536, 2001.
- 9) S. Howard, *Finger talk-South African sign language dictionary*. South Africa: Mondri, 2008.
- 10) Z. Kalal, K. Mikolajczyk and J. Matas, "Forward-backward error: Automatic detection of tracking failures." In *International Conference on Pattern Recognition*, pp. 2756—2759, 2010.
- 11) M. Lee and I. Cohen, "Human upper body pose estimation in static images," in *Proceedings of the European Conference on Computer Vision*, pp. 126—138, 2004.
- 12) Y. Liu and P. Zhang, "Hand gesture tracking using particle filter with multiple features", in *Proceedings of the International Symposium on Intelligent Information Systems and Applications*, pp. 264—267, 2009.
- 13) D. G. Lowe, "Distinctive image features from scale-invariant keypoints", *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91—110, 2004.
- 14) M. Muja and D. G. Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration", in *Proceedings of the International Conference on Computer Vision Theory and Applications*, pp. 331—340, 2009.
- 15) A. Roussos, S. Theodorakis, V. Pitsikalis, and P. Maragos, "Hand tracking and affine shape-appearance handshape subunits in continuous sign language recognition", In *Proceedings of the Workshop on Sign, Gesture and Activity*, 11th European Conference on Computer Vision, pp. 258—272, 2010.
- 16) V. Spruyt, A. Ledda, and S. Geerts, "Real-time multi-colourspace hand segmentation," In *Proceedings of the 17th International Conference on Image processing*, pp. 3117—3120, 2010.

Imran Achmed is a Telkom/Cisco/Aria Technologies/THRIP Centre of Excellence PhD student at the University of the Western Cape. His research focuses on sign language synthesis and novel communication applications for the Deaf and hearing impaired.

Isabella M. Venter is an associate professor and Chair of the Department of Computer Science at the University of the Western Cape.

Peter Eisert is a professor for Visual Computing at the Humboldt University, Berlin, heads the Computer Vision and Graphics Group at the Fraunhofer Heinrich-Hertz Institute and is professor extraordinaire at the University of the Western Cape.